DOCUMENT RESUME

ED 309 199 TM 013 688

AUTHOR Knol, Dirk L.
TITLE Stepwise Item Selection Procedures for Rasch Scales Using Quasi-Loglinear Models. Project Psychometric Aspects of Item Banking No. 44. Research Report 89-3.
INSTITUTION Twente Univ., Enschede (Netherlands). Dept. of Education.
PUB DATE Apr 89
NOTE 35p.
AVAILABLE FROM Bibliotheek, Department of Education, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS *Algorithms; Item Banks; *Latent Trait Theory; Mathematical Models; *Measures (Individuals); *Selection; Simulation; Statistical Analysis; Test Construction; *Test Items
IDENTIFIERS Item Parameters; *Iterative Models; Log Linear Models; Multidimensional Models; Rasch Model; *Rasch Scaled Scores; Unidimensionality (Tests)

ABSTRACT

Two iterative procedures for constructing Rasch scales are presented. A log-likelihood ratio test based on a quasi-loglinear formulation of the Rasch model is given by which one item at a time can be deleted from or added to an initial item set. In the so-called "top-down" algorithm, items are stepwise deleted from a relatively large initial item set, whereas in the "bottom-up" algorithm items are stepwise added to a relatively small initial item set. Both algorithms are evaluated through a simulation study with generated data. Item parameters are given for four generated unidimensional data sets and two generated two-dimensional sets. Abilities were randomly sampled from a multivariate normal distribution with a sample size of 1,000. Results for the top-down algorithm were poor, but results for the bottom-up algorithm were more encouraging. It is suggested that alternating the bottom-up algorithm with one or two iterations of the top-down algorithm would allow the procedure to reject items that were added incorrectly in a previous step. Eight tables illustrate the item parameters and the use of both algorithms for the generated data. (SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED309199

Stepwise Item Selection Procedures for Rasch Scales Using Quasi-Loglinear Models

Research Report 89-3

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

J. NELISSEN

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Dirk L. Knol

department of EDUCATION

University of Twente

Division of Educational Measurement and Data Analysis

TM013688

Project Psychometric Aspects of Item Banking No.44

Colofon:
Typing: L.A.M. Bosch-Padberg
Cover design: Audiovisuele Sectie TOLAB Toegepaste Onderwijskunde
Printed by: Centrale Reproductie-afdeling

Stepwise Item Selection Procedures for Rasch Scales Using Quasi-loglinear Models

Dirk L. Knol

Stepwise item selection procedures for Rasch scales using quasi-loglinear models , Dirk L. Knol - Enschede : University of Twente, Department of Education, April, 1989. - 28 pages

Abstract

Two iterative procedures for constructing Rasch scales are presented. A log-likelihood ratio test based upon a quasi-loglinear formulation of the Rasch model is given by which one item at a time can be deleted from or added to an initial item set. In the so-called top-down algorithm, items are stepwise deleted from a relatively large initial item set whereas in the bottom-up algorithm items are stepwise added to a relatively small initial item set. Both algorithms are evaluated by means of generated data. The results for the top-down algorithm are bad whereas the results for the bottom-up algorithm are more encouraging.

Key words Item selection, Log-likelihood ratio test, Quasi-loglinear models, Rasch model.

Stepwise Item Selection Procedures for Rasch Scales Using Quasi-loglinear Models

When constructing Rasch (1960) scales from a large set of items, it often happens that the Rasch model does not fit to the entire item set. This lack of fit is due to the rather strong assumptions of the Rasch model (cf. Molenaar, 1983), e.g., unidimensionality of the underlying ability and local stochastical independence of item scores. Therefore, usually a two-step procedure is recommended for constructing Rasch scales.

The first step involves the identification of one or more subsets of items approximately satisfying the Rasch model. This identification can, for instance, be based upon a multidimensional representation of the items (cf. Knol, 1986, 1987a) by dividing the space in subspaces.

The second step consists of iteratively deleting one item at a time from a relatively large initial subset or by adding one item at a time to a relatively small initial subset. Usually, deletion of items is based upon item statistics incorporated in computer programs for the Rasch model. For example, the program PML (Molenaar, 1981) gives biserial correlations, U_i statistics (Molenaar, 1983) and contributions of the items to overall goodness of fit tests. However, decisions based upon these indices are highly subjective, partly because of the sometimes contradictory information they provide.

Especially for large scale applications, there is a need for automatic procedures. Such procedures should preferably be based upon sound statistical tests. Some efforts in this direction have already been made. Verhelst (1983) proposed a stepwise procedure based upon a log-likelihood ratio test. However, it seems that his test is statistically not entirely well founded (Knol, 1987b). Moreover, the procedure does not seem to work satisfactory in practice (Knol, 1987b). Another, potentially more promising procedure, comes from quasi-loglinear modeling (Bishop, Fienberg & Holland, 1975; Kelderman, 1987), in which specific hypotheses can be tested. Kelderman (1984) showed that the Rasch model can be written as a quasi-loglinear model. This offers the possibility to detect specific violations of items to the Rasch model.

In this paper, a log-likelihood ratio test based upon a quasi-loglinear formulation of the Rasch model will be presented in which the conditional Rasch model (Fischer, 1974) is tested against an alternative model which incorporates violations to the Rasch model of a particular item. A stepwise top-down procedure based upon this log-likelihood ratio test will be given, in which one item at a time is deleted from a relatively large initial item set. Also, a bottom-up algorithm will be given in which stepwise one item at a time is added to a relatively small initial item set already satisfying the Rasch model. In order to evaluate both algorithms, the procedures will be applied to some generated data sets.

The Rasch Model as a Quasi-loglinear Model

In loglinear models, the logarithms of expected cell frequencies or counts m are explained in terms of linear combinations of functions of observable categorical variables. A subclass of loglinear models arises in the case of a priori or structurally zero cells. These models are called quasi-loglinear models (cf. Bishop, Fienberg & Holland, 1975, Ch. 5). Kelderman (1984, p. 226) showed that the conditional Rasch model (Fischer, 1974) can be written as a quasi-loglinear model. For our purpose it is assumed that the Rasch model contains no subgroups based upon external information such as sex or age. The only subgroups we deal with are score groups. In the analysis of variance or u-terms parametrization (Bishop, Fienberg & Holland, 1975) the logarithms of the expected cell counts m for the conditional Rasch model (without external subgroups) can be written as

$$(1) \qquad \ln m_{x_1 \ldots x_k t} = u + \{ \sum_{j=1}^{k} u_j(x_j)) + u_{k+1}(t) \ .$$

where u is a constant term, $u_j(x_j)$ is the main effect of response x_j (x_j=0,1) of item j (j=1,....,k) and $u_{(k+1)}(t)$ is the main effect of score t (t=0,....,k).

The number of estimable parameters of a quasi-loglinear model is equal to the difference between the number of parameters and the number of constraints imposed by the

model. The number of estimable parameters can be obtained numerically by computing the rank of the so-called design matrix (cf. Bock, 1975, p. 523) of the quasi-loglinear model. An alternative procedure which can be applied for relatively simple models such as (1), consists of counting the number of estimable parameters by correcting for the constraints. This approach will be followed in the present paper. Following the procedure of Kelderman (1984, pp. 231-232) the constant u term counts as one parameter. Furthermore, each term $u_j(x_j)$ $(j=1,\ldots,k)$ counts as one and $u_{(k+1)}(t)$ $(t=0,\ldots,k)$ as $k+1-1=k$ parameters. Finally, we have the constraint $\Sigma_j x_j=t$. Adding the numbers yields the number of estimable parameters of model (1) as $1+k+k-1=2k$. Model (1) can be tested against the fully saturated model by the log-likelihood ratio test or by Pearson's goodness of fit test (Kelderman, 1984). Both test statistics are asymptotically χ^2 distributed with degrees of freedom equal to the difference between the number of structurally nonzero cells and the number of estimable parameters of model (1). However, both test statistics will not be used throughout the paper, since the test statistics are only asymptotically χ^2 distributed and the number of degrees of freedom is 2^k-2k, which is large already for moderate values of k. Instead, model (1) will be tested against a model which allows for each item i $(i=1,\ldots,k)$ separately all first-order interaction terms containing the item response x_i.

For each i $(i=1,\ldots,k)$, the model with all first-order interaction terms containing x_i is

$$(2) \qquad \ln m_{x_1 \ldots x_k t} = u + \{\sum_{j=1}^{k} u_j(x_j)\} + u_{k+1}(t)$$

$$+ \{\sum_{j \neq i} u_{ij}(x_i x_j)\} + u_{i(k+1)}(x_i t)$$

where $u_{ij}(x_i x_j)$ $(i \neq j)$ is the (first-order) interaction between responses x_i and x_j of items i and j and $u_{i(k+1)}(x_i t)$ is the (first-order) interaction between response x_i of item i and score t. An interaction term $u_{ij}(x_i x_j)$ can be interpreted as a measure of local dependence between responses x_i and x_j (Kelderman, 1984, p. 224). Hence, the sum $\Sigma_j u_{ij}(x_i x_j)$ $(j \neq i)$ is a measure of local dependence between item i and the remaining items. The interaction term $u_{i(k+1)}(x_i t)$ can be interpreted as a measure of invariance of the item response function of item i over score groups (Kelderman, 1984, p. 224). In model (2) u-terms are incorporated which reflect both violation of unidimensionality of item i and violations of local independence of that item with the remaining items.

It can be proved that model (2) is separable (cf. Bishop, Fienberg & Holland, Ch. 5) and that the log-likelihood of model (2) equals the sum of the log-likelihoods for model (1), computed separately for the data sets with $x_i=0$ and $x_i=1$. It is easily seen that the number of estimable parameters of model (2) equals twice the number of estimable parameters of model (1) for k-1 items, i.e. $2[2(k-1)]=4k-4$.

Since computation time for model (2) is large compared to model (1), it is more efficient to compute the log-likelihood of model (2) as the sum of the log-likelihoods of model (1) computed separately for the two data sets with $x_i=0$ and $x_i=1$.

For each item i (i=1,...,k), the Rasch model (1) can be tested against model (2) by the log-likelihood ratio test

$$(3) \qquad G_i^2 = -2[L_1 - L_2(i)] = -2[K_1 - K_2(i)] \ ,$$

where L_1 and $L_2(i)$ denote the log-likelihoods of models (1) and (2), respectively, whereas K_1 and $K_2(i)$ are the kernels (Kelderman & Steen, 1988) of the log-likelihoods of models (1) and (2), respectively. Under the assumption of model (1), the test statistic G_i^2 is asymptotically χ^2 distributed with degrees of freedom equal to the difference between the numbers of estimable parameters of model (2) and model (1), i.e. (4k-4)-2k=2k-4. Since computation of the log-likelihood is often impossible and very expensive for larger values of k, (3) will be obtained by computation of the kernels.

In the next section, two algorithms for constructing a Rasch scale will be presented based upon the log-likelihood ratio test (3).

Two Algorithms

Analogous to the algorithm of Verhelst (1983), a top-down algorithm can be constructed in which stepwise one item at a time is deleted from an initial set of items. The computations will be done with the program LOGIMO (Kelderman & Steen, 1988).

Step 1. Start with an initial item set S consisting of (say) k items.

Step 2. Run the program for model (1) for the item set S. Compute for each item $i \epsilon S$ ($i=1,\ldots,k$) the G_i^2 test statistic (3). In our implementation, this involves running the program LOGIMO 2k+1 times. Select the item i^* with the largest G_i^2 value.

Step 3. Compute for the selected item i^* the p-value of the test statistic G_i^2. If $p<.05$ then model (1) is rejected in favour of model (2). This means that unidimensionality of item i^* and/or local independence of that item with the remaining items is violated. If $p<.05$ then delete item i^* from the item set S (i.e. update S) and repeat steps 2 and 3 until no item from set S can be deleted any more.

Step 4. Evaluate the constructed scale with Andersen's (1973) log-likelihood ratio test and with the Martin-Löf (1973) χ^2 test.

It is also possible to define a bottom-up algorithm (cf. Verhelst, 1983), in which stepwise an item is added to a (small) set of items already satisfying the Rasch model. A bottom-up algorithm based upon the log-likelihood ratio test (3) is statistically more appropriate than the top-down algorithm, because the test statistic (3) is only χ^2 distributed when the null-hypothesis (i.e. model (1)) is true (cf. Verhelst, 1983). For the top-down algorithm the assumption that the Rasch model (1) holds, can hardly be made. However, for the bottom-up algorithm this assumption can be made, provided that it is possible to select a (small) initial item set that satisfies the Rasch model.

A bottom-up algorithm based upon the log-likelihood ratio test (3) can be stated as:

Step 1. Start with an initial set S' of k' items that already satisfies the Rasch model and a non-overlapping set C of n items containing the items that can potentially be added to the Rasch scale.

Step 2. Compute for each item $i \in C$ $(i=1,\ldots,n)$ the G_i^2 test statistic (3) for the k'+1 items of the set S'+{i}. In our implementation, this involves running the program LOGIMO 3n times. Select the item i^* with the smallest G_i^2 value.

Step 3. Compute for the selected item i^* the p-value of G_i^2. If $p>.05$ then model (1) cannot be rejected. This means that the set of items S+{i^*) satisfies the Rasch model. If $p>.05$ then add item i^* to the item

set S (i.e. update S' and C) and repeat steps 2 and 3 until no item from set C can be added any more.

Step 4. Evaluate the constructed scale with Andersen's (1973) log-likelihood ratio test and with the Martin-Löf (1973) χ^2 test.

For each iteration cycle of the bottom-up algorithm, the program has to be run 3n times whereas the number of runs for the top-down algoritm is only 2k+1. However, since the cardinality k' of the start set S' of the bottom-up algorithm is typically much smaller (especially during the first iteration cycles) than the cardinality k of the start set S of the top-down algorithm, it can be expected that CPU-time for the bottom-up algorithm will be less than that for the top-down algorithm.

In the next section the performances of the top-down and bottom-up algorithms will be evaluated empirically using some generated data sets.

A Simulation Study

Data have been generated according to the two-parameter multidimensional logistic model (Reckase, 1973). In this model, the item response function for item i is given by

$$(4) \qquad p_i(\underline{\theta}) = (1 + \exp[-1.6(\underline{\alpha}_i'\underline{\theta} - \beta_i)])^{-1} ,$$

where $\underline{\alpha}_i=(\alpha_{i1},\ldots,\alpha_{im})'$ is a vector of item discrimination parameters, m is the dimensionality of the ability space, β_i is the item difficulty parameter and $\underline{\theta}=(\theta_1,\ldots,\theta_m)$ is the m-dimensional vector of abilities. Note that model (4) allows for items with different dimensionality. For example, with (4) items can be generated which have a nonzero discrimination parameter on one dimension and zeroes on the remaining dimensions. Note also that (4) reduces to the Rasch model if m=1 and $\alpha_i=\alpha$ (constant) for all items i. Model (4) allows for items violating the Rasch model in the sense of different slopes (discrimination) and different dimensionality (violating the local independence). In Table 1 the item parameters of four generated unidimensional data sets are given, and in Table 2 the item parameters of two generated two-dimensional data sets.

Insert Tables 1 and 2 about here

The data sets are constructed such that the items 1 through 10 of each data set form a Rasch scale with discrimination parameters $\alpha_i=\alpha=1$ (i=1,...,10). In all data sets the items 11 through 15 differ in discrimination (sets 1 through 4 and 6) and/or dimensionality (sets 5 and 6) from the (dominant) Rasch scale.

Abilities were randomly sampled from a multivariate normal (0,I) distribution. The sample size was chosen to be 1000, which was expected to be large enough to get sufficiently reliable results. It is likely that the repeated use of the test (3) in the algorithms will result in chance capitalization. For each data set a second, independent data set with sample size 1000 was generated to evaluate the final scales found by the algorithms.

Both the top-down and the bottom-up algorithm were applied to all data sets. The top-down algorithm starts with the item set S, consisting of all the 15 items, whereas the bottom-up algorithm has the startset S', consisting of the items 4 through 7, and the remaining items form the set C of candidate items.

Since in model (2) only first-order interaction terms have been incorporated and no overall goodness of fit test is available for model (2) because of the too large number of degrees of freedom, it is necessary to evaluate the item selection procedure by an external criterion. In the program PML (Molenaar, 1981) the χ^2 goodness of fit test of Martin-Löf (1973) and the log-likelihood ratio test of Andersen (1973) are implemented. Both tests were used to evaluate the obtained Rasch scale after the selection procedures.

From Table 1 it can be seen that data set 1 contains, besides the dominating Rasch scale (items 1 through 10) a subscale consisting of the items 11 through 15 which have higher discrimination parameters (α=1.4) than the discrimination parameters of the dominating scale (α=1). Data

set 2 contains a subscale consisting of relatively low discriminations (α=.6). The results of the algorithms for the data sets 1 and 2 are given in Tables 3 and 4 , respectively.

Insert Tables 3 and 4 about here

In order to give more insight in the item selection procedures, the p-values of Martin-Löf's χ^2 and Andersen's log-likelihood ratio test are given after each added or deleted item. Additionally, as a baseline the p-values of both tests are given for the theoretically expected scale(s). Also, the p-values of both tests are given for the cross-validated scales. From Tables 3 and 4 it is seen that the top-down algorithm clearly fails to detect the dominating scale. In both cases, the obtained scale consists of a mixture of items of the dominating scale and the subscale. The outcomes for the data sets 1 and 2 of the bottom-up algorithm are better: the obtained scales contain all items of the dominating scale and (only) two items of the subscale. In a sense, the bottom-up algorithm iterates too long. After k=9, the algorithm starts to select items from the subscale.

Contrary to the first two data sets, the data sets 3 and 4 contain no substantial subscale. The difference between data set 3 and 4 is that the discrimination parameters of the former are more extreme than those of the latter. The results

of the algorithms for data sets 3 and 4 are given in Tables 5 and 6, respectively.

Insert Tables 5 and 6 about here

From Table 5 it can be seen that the top-down algorithm for data set 3 yields the dominating scale. The top-down algorithm applied to data set 4, however, yields a mixture of items from the dominating scale and three of the remaining items. Obviously, the algorithm cannot differentiate well between items of the dominating scale with discrimination parameter $\alpha=1$ and items with a slightly different discrimination parameter ($\alpha=.8$ or $\alpha=1.2$). However, the resulting scale has good overall goodness of fit values. As in the case of data sets 1 and 2, the bottum-up algorithm for data sets 3 and 4 iterates too long. After an optimum has been reached ($k=9$ for data set 3 and $k=10$ for data set 4), the algorithm still adds items whereas it would be better if the algorithm had been stopped.

The data sets 5 and 6 are two-dimensional, where the items 11 through 15 measure another trait than the items 1 through 10 do. Data set 5 contains a Rasch subscale consisting of the items 11 through 15 whereas data set 6 does not. The results of the algorithms for data sets 5 and 6 are given in the Tables 7 and 8, respectively.

Insert Tables 7 and 8 about here

From the Tables 7 and 8 it can be seen that in both cases the top–down algorithm clearly fails to yield the dominating Rasch scale. For data set 5, the top–down algorithm could not discriminate well between the two subscales. Firstly, the items 3, 4, 5, 7 and 8 of the dominating scale with moderate difficulty parameters are deleted. Secondly, all items of the subscale (items 11 through 15) are deleted. However, the resulting scale has good overall goodness of fit values and contains only items of the dominating scale. Nevertheless the outcome is not satisfactory because the scale consists of only five items. In both cases the bottom–up algorithm performs well, yielding in both cases the dominating scale consisting of the items 1 through 10.

Summarized, the top–down algorithm yields the complete dominating scale in only in one case (data set 3). The performance of the bottom–up algorithm is more encouraging. For the data sets 5 and 6 where the items 11 through 15 measure another trait than the trait measured by the items of the dominating scale, the algorithm yields precisely the dominating scale. For the other data sets where the items 11 through 15 only differ in discrimination parameters from the items of the dominating scale, the resulting scale consists of all items of the dominating scale (except for data set 4) and only one or two items of the second subscale. For data

set 4, the items 1 and 2 do not belong to the resulting scale.

Discussion

From the results presented in the last section, it is clear that the bottom-up algorithm is more promising than the top-down algorithm. Moreover, the bottom-up algorithm is statistically better justified than the top-down algorithm (cf. Verhelst, 1983). Finally, the bottom-up algorithm is typically faster than the top-down algorithm because the former starts with a smaller initial item set. However, it has to be noted that CPU-times for the bottom-up algorithm are still very large.

Furthermore, in the presented simulation study the bottom-up algorithm starts with an initial item set already forming a Rasch scale. Of course, in practice we do not have this knowledge. Without a priori knowledge, it seems very difficult to select a small item set that satisfies the Rasch model.

In the simulation study it seems that the bottom-up algorithm iterates too long: after cycle 4 or 5 the algorithm starts to select items from the subscale. An improvement could be to increase the significance level of the log-likelihood ratio test (3). Another, probably better possibility is to alternate the bottom-up algorithm with one or two iterations of the top-down algorithm. This allows the

procedure to reject items that have been added incorrectly to the scale in a previous step. An additional advantage of such a mixed procedure would be that the choice of the startset is less critical.

In the alternative model (2), only first-order interaction terms have been incorporated. A possible explanation of the rather disappointing outcomes of both algorithms can be that higher-order interaction terms are needed to describe the (induced) violations against the Rasch model. However, incorporating higher-order interactions in the alternative model will make the algorithms much more expensive and, even worse, it will be impossible to run the top-down algorithm for large item sets. Finally, it has to be noted that because of the repeated use of the test (3), it is likely that chance capitalization occurs. Therefore, with real data the final scales have to be cross-validated in an independent sample.

References

Andersen, E.B. (1973). A goodness of fit test for the Rasch model. Psychometrika, 38, 123–140.

Bishop, Y.M.M, Fienberg, S.E. & Holland, P.W. (1975). Discrete multivariate analysis. Cambridge, MA: The MIT Press.

Bock, R.D. (1975). Multivariate statistical methods in behavioral research. New York: McGraw–Hill.

Fischer, G.H. (1974). Einführung in die Theorie psychologischer Tests. Bern: Huber.

Gustafsson, J.E. (1980). Testing and obtaining fit of data to the Rasch model. British Journal of Mathematical and Statistical Psychology, 33, 205–233.

Kelderman, H. (1984). Loglinear Rasch model tests. Psychometrika, 49, 223–245.

Kelderman, H. (1987). Quasi–loglinear models for test and item analysis. Unpublished doctoral dissertation. Enschede, The Netherlands: University of Twente.

Kelderman, H. & Steen, R. (1988). Logimo I. Loglinear item response theory modeling (computer manual). Enschede, The Netherlands: University of Twente, Department of Education.

Knol, D.L. (1986). Inventarisatie van automatische itemselectie procedures voor Raschschalen (R–86–2). Enschede: Universiteit Twente.

Knol, D.L. (1987a). Het verband tussen item respons theorie en factoranalyse voor dichotome items (R-87-2). Enschede: Universiteit Twente.

Knol, D.L. (1987b). Stapsgewijze itemselectieprocedures in het Raschmodel (R-87-6). Enschede: Universiteit Twente.

Martin-Löf, P. (1973). Statistika modeller. Antecknigar från seminarier läsaret 1969-70 utarbetade av Rolf Sundberg, 2:a uppl. Stockholm: Institute för försäkringsmatematik och matematik statistik vid Stockholms universitet.

Molenaar, I.W. (1981). Programmabeschrijving van PML (versie 3.1) voor het Rasch-model (HB-81-538-RP). Groningen: Rijksuniversiteit Groningen, Vakgroep Statistiek en Meettheorie, FSW.

Molenaar, I.W. (1983). Some improved diagnostics for failure of the Rasch model. Psychometrika, 48, 49-73.

Rasch, G. (1960). Probabilistic models for some intelligence and attainment tests. Copenhagen: The Danish Institute for Educational Research.

Reckase, M.D. (1973). Development and application of a multivariate logistic latent trait model. Dissertation Abstracts International, 33.

Verhelst, N. (1983). Automatische itemselectieprocedures in het Rasch-model. Utrecht: Rijksuniversiteit Utrecht. Subfaculteit Psychologie, Vakgroep P.S.M.

Table 1
Itemparameters of the four generated unidimensional data sets 1 through 4.

	set							
	1		2		3		4	
item	α	β	α	β	α	β	α	β
1	1	−1.8	1	−1.8	1	−1.8	1	−1.8
2	1	−1.4	1	−1.4	1	−1.4	1	−1.4
3	1	−1.0	1	−1.0	1	−1.0	1	−1.0
4	1	−0.6	1	−0.6	1	−0.6	1	−0.6
5	1	−0.2	1	−0.2	1	−0.2	1	−0.2
6	1	0.2	1	0.2	1	0.2	1	0.2
7	1	0.6	1	0.6	1	0.6	1	0.6
8	1	1.0	1	1.0	1	1.0	1	1.0
9	1	1.4	1	1.4	1	1.4	1	1.4
10	1	1.8	1	1.8	1	1.8	1	1.8
11	1.4	−2.0	0.6	−2.0	0.6	−1.0	0.8	−1.0
12	1.4	−1.0	0.6	−1.0	0.6	1.0	0.8	1.0
13	1.4	0.0	0.6	0.0	1.4	−1.0	1.2	−1.0
14	1.4	1.0	0.6	1.0	1.4	0.0	1.2	0.0
15	1.4	2.0	0.6	2.0	1.4	1.0	1.2	1.0

Table 2
Itemparameters of the two generated two-dimensional data sets 5 and 6.

	set					
	5			6		
item	α_1	α_2	β	α_1	α_2	β
1	1	0	-1.8	1	0	-1.8
2	1	0	-1.4	1	0	-1.4
3	1	0	-1.0	1	0	-1.0
4	1	0	-0.6	1	0	-0.6
5	1	0	-0.2	1	0	-0.2
6	1	0	0.2	1	0	0.2
7	1	0	0.6	1	0	0.6
8	1	0	1.0	1	0	1.0
9	1	0	1.4	1	0	1.4
10	1	0	1.8	1	0	1.8
11	0	1	-2.0	0	0.6	-1.0
12	0	1	-1.0	0	0.6	1.0
13	0	1	0.0	0	1.4	-1.0
14	0	1	1.0	0	1.4	0.0
15	0	1	2.0	0	1.4	1.0

Table 3

P-values of the Martin-Löf χ^2 and Andersen's log-likelihood ratio test of the top-down and bottom-up algorithms for data set 1.

algorithm	scale	k	added/ deleted item	test statistic χ^2	test statistic LR
baseline	1-10	10		.36	.31
	11-15	5		.66	—*
top-down	1-15	15		.27	.00
		14	-6	.13	.00
		13	-5	.23	.01
		12	-7	.47	.00
	1-4,8-11,13-15	11	-12	.91	—*
cross-validation		11		.00	.12
bottom-up	4-7	4		.02	.00
		5	+3	.05	.96
		6	+10	.40	.10
		7	+2	.89	.13
		8	+8	.80	.04
		9	+1	.79	.61
		10	+15	.84	.00
		11	+12	.09	.00
	1-10,12,15	12	+9	.03	.00
cross-validation		12		.00	.14

* Andersen's LR test cannot be computed.

Table 4

P-values of the Martin-Löf χ^2 and Andersen's log-likelihood ratio test of the top-down and bottom-up algorithms for data set 2.

algorithm	scale	k	added/ deleted item	test statistic χ^2	test statistic LR
baseline	1-10	10		.61	.83
	11-15	5		.70	96
top-down	1-15	15		.00	.00
		14	-2	.00	.00
		13	-5	.01	.01
		12	-7	.04	.00
		11	-4	.24	.00
		10	-6	.34	—*
	1,8-15	9	-3	.87	—*
cross-validation		9		.10	.03
bottom-up	4-7	4		.99	.96
		5	+9	.95	.81
		6	+10	.95	.73
		7	+3	.92	.63
		8	+1	.95	.68
		9	+2	.88	.45
		10	+15	.93	.77
		11	+8	.81	.24
	1-11,15	12	+11	.89	.09
cross-validation		12		.14	.01

* Andersen's LR test cannot be computed.

Table 5

P-values of the Martin-Löf χ^2 and Andersen's log-likelihood ratio test of the top-down and bottom-up algorithms for data set 3.

algorithm	scale	k	added/ deleted item	test statistic χ^2	test statistic LR
baseline	1-10	10		.94	.46
top-down	1-15	15		.00	.00
		14	-14	.00	.00
		13	-12	.01	.00
		12	-13	.06	.03
		11	-11	.69	.64
	1-10	10	-15	.94	.46
cross-validation		10		.57	.24
bottom-up	4-7	4		.78	.91
		5	+8	.75	.91
		6	+3	.95	.98
		7	+1	.99	.83
		8	+2	.99	.50
		9	+10	.99	.73
		10	+14	.35	.57
	1-10,14	11	+9	.18	.29
cross-validation		11		.33	.05

Table 6

P-values of the Martin-Löf χ^2 and Andersen's log-likelihood ratio test of the top-down and bottom-up algorithms for data set 4.

algorithm	scale	k	added/ deleted item	test statistic χ^2	test statistic LR
baseline	1-10	10		.48	.67
top-down	1-15	15		.16	.06
		14	-7	.17	.01
		13	-3	.25	.02
		12	-15	.17	.06
		11	-12	.28	.06
		10	-4	.24	.07
	1,2,5,8-11,13,14	9	-6	.87	.85
cross-validation		9		.88	.14
bottom-up	4-7	4		.14	.82
		5	+9	.38	.95
		6	+11	.92	.95
		7	+3	.78	.78
		8	+15	.71	.39
		9	+8	.66	.33
		10	+10	.75	.78
	3-11,13,15	11	+13	.31	.02
cross-validation		11		.64	.63

Table 7

P-values of the Martin-Löf χ^2 and Andersen's log-likelihood ratio test of the top-down and bottom-up algorithms for data set 5.

algorithm	scale	k	added/ deleted item	test statistic χ^2	test statistic LR
baseline	1-10	10		.49	.42
	11-15	5		.27	—*
top-down	1-15	15		.00	.00
		14	-8	.00	.00
		13	-4	.00	.00
		12	-7	.00	.00
		11	-3	.04	.00
		10	-5	.04	.00
		9	-12	.09	.00
		8	-13	.00	.00
		7	-14	.00	.00
		6	-15	.00	.00
	1,2,6,9,10	5	-11	.87	.68
cross-validation		5		.75	—*
bottom-up	4-7	4		.46	.29
		5	+3	.95	.55
		6	+8	.76	.44
		7	+1	.78	.44
		8	+9	.35	.52
		9	+10	.22	.24
	1-10	10	+2	.49	.42
cross-validation		10		.75	.60

* Andersen's LR test cannot be computed.

Table 8

P-values of the Martin-Löf χ^2 and Andersen's log-likelihood ratio test of the top-down and bottom-up algorithms for data set 6.

algorithm	scale	k	added/ deleted item	test statistic χ^2	test statistic LR
baseline	1-10	10		.31	.70
top-down	1-15	15		.00	.00
		14	-15	.00	.00
		13	-14	.00	.00
		12	-7	.00	.00
		11	-4	.00	.00
		10	-6	.00	.00
		9	-5	.00	.00
		8	-3	.00	.05
		7	-9	.67	.40
		6	-2	.01	.35
		5	-13	.00	.08
		4	-8	.31	.87
	1,11,12	3	-10	.09	.10
cross-validation		3		.63	.56
bottom-up	4-7	4		.65	.73
		5	+8	.63	.69
		6	+2	.53	.54
		7	+1	.41	.29
		8	+10	.26	.37
		9	+9	.49	.53
	1-10	10	+3	.31	.70
cross-validation		10		.77	.63

Titles of recent Research Reports from the Division of Educational Measurement and Data Analysis, University of Twente, Enschede, The Netherlands.

RR-88-1 E. van der Burg & J. de Leeuw, *Nonlinear redundancy analysis*

RR-88-2 W.J. van der Linden & J.J. Adema, *Algorithmic test design using classical item parameters*

RR-88-3 E. Boekkooi-Timminga, *A cluster-based method for test construction*

RR-88-4 J.J. Adema, *A note on solving large-scale zero-one programming problems*

RR-88-5 W.J. van der Linden, *Optimizing incomplete sample designs for item response model parameters*

RR-88-6 H.J. Vos, *The use of decision theory in the Minnesota Adaptive Instructional System*

RR-88-7 J.H.A.N. Rikers, *Towards an authoring system for item construction*

RR-88-8 R.J.H. Engelen, W.J. van der Linden, & S.J. Oosterloo, *Item information in the Rasch model*

RR-88-9 W.J. van der Linden & T.J.H.M. Eggen, *The Rasch model as a model for paired comparisons with an individual tie parameter*

RR-88-10 H. Kelderman & G. Macready, *Loglinear-latent-class models for detecting item bias*

RR-88-11 D.L. Knol & M.P.F. Berger, *Empirical comparison between factor analysis and item response models*

RR-88-12 E. van der Burg & G. Dijksterhuis, *Nonlinear canonical correlation analysis of multiway data*

RR-88-13 J. Kogut, *Asymptotic distribution of an IRT person fit index*

RR-88-14 J.J. Adema, *The construction of two-stage tests*

RR-88-15 H.J. Vos, *Simultaneous optimization of decisions using a linear utility function*

RR-88-16 H. Kelderman, *An IRT model for item responses that are subject to omission and/or intrusion errors*

RR-88-17 H. Kelderman, *Loglinear multidimensional IRT models for polytomously scored items*

RR-88-18 H.J. Vos, *Applications of decision theory to computer based adaptive instructional systems*

RR-89-1 R.J.H. Engelen & R.J. Jannarone, *A connection between item/subtest regression and the Rasch model*

RR-89-2 E. Boekkooi-Timminga, *The construction of parallel tests from IRT-based item banks*

RR-89-3 D.L. Knol, *Stepwise item selection procedures for Rasch scales using quasi-loglinear models*

Research Reports can be obtained at costs from Bibliotheek, Department of Education, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.

department of
EDUCATION

A publication by
the Department of Education
of the University of Twente
P.O. Box 217
7500 AE Enschede
The Netherlands